High-Throughput Sequencing of Transposable Elements Insertions Provides Evidence for Adaptive Evolution of the Invasive Asian Tiger Mosquito Towards Temperate Environments

Clément Goubert[1], Hélène Henri[1], Guillaume Minard[2,3], Claire Valiente Moro[2], Patrick Mavingui[2,4], Cristina Vieira[1] and Matthieu Boulesteix[1]

[1]Université de Lyon, F-69622, Lyon, France; Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, F-69100 Villeurbanne, France

[2]Université de Lyon, F-69622, Lyon, France; Université Lyon 1, Villeurbanne, France; CNRS, UMR 5557, Ecologie Microbienne, Villeurbanne, France; INRA, UMR 1418, Villeurbanne, France

[3]Metapopulation Research Center, Department of Biosciences, University of Helsinki, Helsinki, Finland

[4]Université de La Réunion, UMR PIMIT, INSERM 1187, CNRS 9192, IRD 249, Plateforme Technologique CYROI, Sainte-Clotilde, La Réunion.

Keywords: Invasive species, *Aedes albopictus*, local adaptation, genome scan, Transposable Elements, diapause

Corresponding author: Matthieu Boulesteix, Université de Lyon, F-69622, Lyon, France; Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, F-69100 Villeurbanne, France; phone: +334 72 43 29 16, fax:

20 +334 72 43 13 88; email: matthieu.boulesteix@univ-lyon1.fr

21 **Running title:** Molecular adaptation in *Ae. albopictus*

Abstract

Invasive species represent unique opportunities to evaluate the role of local adaptation during colonization of new environments. Among these, the Asian tiger mosquito, *Aedes albopictus*, is a threatening vector of several human viral diseases, including dengue, chikungunya and the emerging Zika fevers. Its broad presence in both temperate and tropical environments has sometimes been considered as the reflect of a great "ecological plasticity". However, no study has been conducted to assess the role of adaptive evolution in the ecological success of *Ae. albopictu*s at the molecular level. In the present study we performed a genomic scan to search for potential signatures of selection leading to local adaptation in a hundred of field collected mosquitoes from native populations of Vietnam and temperate invasive populations of Europe. High throughput genotyping of transposable element insertions generated more than 120 000 polymorphic loci, which in their great majority revealed a virtual absence of structure between bio-geographic areas. Nevertheless, 92 outlier loci show a high level of differentiation between temperate and tropical populations. The majority of these loci segregates at high insertion frequencies among European populations, indicating that this pattern could have been caused by recent events of adaptive evolution in temperate areas. Six outliers were located near putative diapause effector genes, suggesting fine-tuning of this critical pathway during

42 local adaptation.

Introduction

Biological invasions represent unique opportunities to study fast evolutionary changes such as adaptive evolution. Indeed, settlement in a novel area represents a biological challenge that invasive species have successfully overcome. The underlying processes could be studied at the molecular level, particularly to gather empirical knowledge about the genetics of invasions, a field of study that has produced extensive theoretical predictions, but for which there is still little evidence in nature (Colautti & Lau 2015). Some of the main concerns are to disentangle the effects of neutral processes during colonization, such as founder events or allele surfing at the migration front, from adaptive evolution (i.e. local adaptation, Lande 2015; Peischl & Excoffier 2015; Colautti & Lau 2015).

Adaptation can arise either through the appearance and spread of a new beneficial mutation, the spread of a favorable allele from standing genetic variation, or from hybridization in the introduction area (Handley *et al.* 2011; Colautti & Lau 2015; Bock *et al.* 2015). Detection of the footprint of natural selection is however dependent on the availability of informative genetic markers, which should provide a substantial coverage of the genome to allow selection scans and be easily and confidently scored across many individuals. Unfortunately, invasive organisms are rarely model species, making the development of a reliable and efficient marker challenging.

The Asian tiger mosquito, *Aedes* (*Stegomya*) *albopictus* (Diptera:Culicidae) is currently one of the most threatening invasive species (Invasive Species Specialist Group).

Originating from South-Eastern Asia, this species is one of the primary vectors of Dengue and Chikungunya viruses, and is also involved in the transmission of other threatening arboviruses (Paupy *et al.* 2009), in particular the newly emerging Zika virus (Grard *et al.* 2014; Marcondes & Ximenes 2015; Chouin-Carneiro *et al.* 2016). *Ae. albopictus* has now settled in every continent except Antarctica, and is found both under tropical and temperate climates (Bonizzoni *et al.* 2013). While this species is supposed to originate from rain forests of South-Eastern Asia (Hawley 1988), the native area of *Ae. albopictus* encompasses contrasted environments including temperate regions of Japan and China, offering a large potential of fit towards newly colonized environments. For example, the induction of photoperiodic diapause in temperate areas, that has a genetic basis in *Ae. albopictus* (Hawley *et al.* 1987; Hanson & Craig 1994; Urbanski *et al.* 2010), is decisive to ensure invasive success in Europe or Northern America. Indeed it allows the susceptible populations to survive through winter at the larval stage into the eggs. Such a trait appears governed by a “genetic toolkit” involving numerous genes and metabolic networks, for which however the genetic polymorphism between diapausing and non-diapausing strains remains to be elucidated (Poelchau *et al.* 2013b). In addition, the colonization of new areas that look similar at first glance can still involve *de novo* adaptation: indeed, even environment sharing climatic variables are not necessarily similar regarding edaphic and biotic interactions (Colautti & Lau 2015). Hence, this suggests that whatever are the native and settled environments, it might be possible to find evidence of adaptive evolution in invasive populations of *Ae. albopictus*.

To better understand the invasive success of this species, we genotyped 140 field

individuals, collected from three Vietnamese (native tropical area) and five European (invasive temperate area) populations, aiming to identify genomic regions involved in local adaptation. To do so, we developed new genetic markers, based on high throughput genotyping of the insertion of Transposable Elements (TEs) that represent at least one third of the genome of *Ae. albopictus* and include recently active families that could reach thousands of copies in one genome (Goubert *et al.* 2015).

Amplification of TE insertions is particularly efficient to obtain a large number of genetic markers throughout one genome (Bonin *et al.* 2008), especially if few genomic resources are available (Monden *et al.* 2014), which was case until recently for the Asian tiger mosquito. In addition, such markers represent a seducing alternative to other methods of diversity reduction, such as RAD-sequencing (Miller et al. 2007), that could be less efficient in species with high TE load (Davey et al. 2012) and did not produce satisfying results in *Ae. albopictus* (Goubert *et al.* 2016). In mosquitoes, TEs have been shown to be powerful markers both for population structure analysis (Biedler *et al.* 2003; Boulesteix *et al.* 2007; Santolamazza *et al.* 2008; Esnault *et al.* 2008) and genome scans (Bonin *et al.* 2009). We hypothesized that some TE insertion sites could be located at the neighborhood of targets of natural selection and thus could reach high level of differentiation between native and invasive populations if selective sweeps occurred during local adaptation. In addition, some TEs could also insert near or inside coding regions and thus could be directly involved in environmental adaptation (Casacuberta & González 2013), eventually contributing to the success of invasive species (Stapley *et al.* 2015).

To distinguish between neutral demographic effects and adaptive evolution, we first performed population genetic analyses to reveal the global genetic structure of the studied populations. We then performed a genomic scan for selection and identified 92 candidate loci under directional selection, among which several can be located in the neighborhood of diapause related genes.

Material and Methods

Biological samples

A total number of 140 flying adult females *Ae. albopictus* were collected in the field at eight sampling sites in Europe and Vietnam during the summers of 2012 and 2013 (Figure 1 and Table S1). Individuals were either sampled using a single trap or using aspirators through the sampling site within a 50 meters radius. When traps were used, live mosquitoes were collected after a maximum of two days.

High throughput Transposon Display (TD) genotyping.

Insertion polymorphism of five transposable elements families: I Loner Ele1 (IL1), Loa Ele2B (L2B), RTE4, RTE5 and Lian 1 identified by Goubert *et al.* (2015) in *Ae. albopictus* was characterized using Transposon Display (TD), a TE insertion specific PCR method, combined with Illumina sequencing of all TD amplification products (Figure S1). These TE families were chosen according to their high estimate of copy number (from 513 to 4203 copies), high identity between copies, and a “copy and paste” mode of transposition (all these TEs are non-LTR Class I retrotransposons).

DNA Extraction and TD adapter ligation. Total DNA was extracted from whole adult bodies following the Phenol-Chloroform protocol described by Minard *et al.* (2015). TD was then set up combining methods from previous studies (Munroe *et al.* 1994; Roy *et al.* 1999; Akkouche *et al.* 2012; Carnelossi *et al.* 2014). First, individual extracted DNA (≈ 75ng) was used for enzymatic digestion in a total volume of 20 *µ*L, with HindIII enzyme (10U/*µ*L) and buffer R (Thermo Scientific) for 3 hours at 37C. The enzyme was inactivated at 80C for 20 minutes. TD adapters were then build hybridizing Hindlink with MSEB oligonucleotides (100*µ*M, see Table S2) in 20X SSC and 1M Tris in a total volume of 333*µ*L after 5mn of initial denaturation at 92C and 1h at room temperature for hybridization of the two parts. Once ready, TD adapters were then ligated to 20 *µ*L of the digested DNA mixing 2*µ*L of TD adapter with 10U T4 ligase and 5X buffer (Fermentas) in a final volume of 50*µ*L for 3 hours at 23C.

Library construction. For each individual, and for each of the five TE families, TE insertions were amplified by PCR (PCR 1) in a Biorad Thermal Cycler (either C1000 or S1000), in a final volume of 25*µ*L. Mixture contained 2*µ*L of digested-ligated DNA with 1*µ*L dNTPs (10mM), 0.5*µ*L TD-adapter specific primer (LNP, 10*µ*M, see S3 Table) and 0.5 *µ*L of TE specific primer (10*µ*M), 1U AccuTaq polymerase (5U/*µ*L) with 10X buffer and Dimethyl-Sulfoxyde (Sigma). Amplification was performed as follows: denaturation at 98C for 30 seconds then 30 cycles including 94C for 15 seconds, hybridization at 60C for 20 seconds and elongation at 68C for 1 minute; final elongation was performed for 5 minutes at 68C. For L2B and RTE5 TEs, a nested PCR was performed in order to

increase specificity in the same PCR conditions using internal forward TE primers and LNP (Table S2). PCR 1 primers include a shared tag sequence that was used for hybridization of the individual indexes by PCR 2.

For each TE, three independents PCR 1 were performed from the same digestion product. PCR 1 products (3 PCR * 5 TE per individual) were then purified using volume-to-volume Agencout AMPure XP beads (20μL PCR 1 + 20μL beads) and eluted in 30μL Resuspension buffer. After nanodrop quantification, equimolar pools containing the 3*5 PCR products per individual were made using Tecan EVO200 robot. Individual pools were then size selected for fragment ranging from 300 to 600 bp using Agencout AMPure XP beads as follow: first magnetic beads were diluted in H2O with a ratio of 1:0.68 then add to 0.625 X PCR products in order to exclude long fragments. A second purification was performed using a non-diluted bead: DNA ratio of 1:8.3 to exclude small fragments.

Samples' multiplexing was performed using home made 6 bp index (included in SRA individual name), which were added to the R primer (Table S2) during a second PCR (PCR 2) with 12 cycles in ABI 2720 Thermal Cycler. Mixture contained 15ng PCR products, 1μl of dNTPs (10mM), 0.5μl MTP Taq DNA Polymerase (5U/μl, Sigma), 5μl 10X MTP Taq Buffer and 1.25μl of each tagged-primer (20μm) in a final volume of 50μl. Amplification was performed as follows: denaturation at 94C for 60 seconds then 12 cycles including denaturation at 94C for 60 seconds, hybridization at 65C for 60 seconds and elongation at 72C for 60 seconds; final elongation was performed for 10 minutes at 72C. PCR 2 products were purified using Agencout AMPure XP beads: DNA ratio of

1:1.25 to obtain libraries. Finally, TD products were paired-end sequenced on an Illumina Hiseq 2000 (1 lane) at the GeT-PlaGe core facility (Genome and Transcriptome, Toulouse) using TruSeq PE Cluster Kit v3 (2x100 bp) and TruSeq SBS Kit v3.

Bioinformatic treatment of TD sequencing.

The different steps of the informatics treatment from the raw sequencing dataset to population binary matrices for presence/absence of TE insertions per individual are described in Figure S2. First, the paired-end reads of each individual were quality checked and trimmed using UrQt v. 1.0.17 (Modolo & Lerat 2015) with standard parameters and a *t* quality threshold of 10. Reads pairs were then checked and trimmed for Illumina adapter contamination using cutadapt (Martin 2011). Specific amplification of TE insertions was controlled by checking for the expected 3' TE sequence on the R1 read using Blat (Altschul *et al.* 1990) with an identity threshold of 0.90. Only reads with an alignment-length/read-length ratio ≥ 0.90 were then retained. R2 reads for which the R1 mate passed this filter were then selected for the insertion loci construction, after the removal of the TD adapter on the 5’ start using cutadapt and the removal of reads under 30 bp. Selected reads were separated in each individual according to the TE families for loci construction.

In order to correct the inter-individual coverage variations, we performed a sampling of the cleaned reads. First, for each TE family, distribution of the number of read per individual was drown, and individuals with less reads than the first decile of this distribution were removed; then cleaned reads of the remaining individuals were

194 randomly sampled at the value of the first decile of coverage (this value varies among
195 TE families). For each TE, the sampled reads of each retained individual were clustered
196 together using the CD-HIT-EST program (Li & Godzik 2006) to recover insertion loci.
197 During this all-to-all reads comparison, the alignments must had a minimum of 90
198 percent identity, and the shortest sequence should be 95% length of the longest, global
199 identity was used and each read was assigned at its best cluster (instead of the first that
200 meet the threshold). In a second step, the reference reads of each locus within
201 individual, given by CD-HIT-EST, were clustered with all the reference reads of all
202 individuals, using the same threshold, in order to build the locus catalog including list of
203 loci of all individuals and the coverage for each locus in each individual. After this step,
204 insertion loci that matched known repeats of the Asian tiger mosquito (Goubert *et al.*
205 2015) were discarded; alignments were performed with Blastn using default parameters.

206 Since the quality control removed a substantial number of reads for the construction of
207 TE insertions catalog, the raw R2 reads (with their TD adapter removed), that could
208 have been discarded in a first attempt were then mapped over the catalog in order to
209 increase the scoring sensibility. Before mapping, the raw R2 reads were also sampled at
210 the first decile of individual coverage (as described previously). At this step, individuals
211 that have been removed from at least two TE families for loci construction were
212 definitively removed from the whole analysis. Mapping was performed over all insertion
213 loci of all TE families in a single run in order to prevent multiple hits. Blat was used with
214 an identity threshold of 90 percent. Visual inspection of alignment quality over 30
215 sampled loci per TE family was performed in order to ensure the quality of scoring.

In order to check if the sampling procedure would affect our results, the read sampling procedures and subsequent analysis were performed independently 3 times (replicates M1, M2 and M3).

Genetic analyses and Genomic scan.

Population structure analyses were performed independently for each TE family. Principal Coordinate Analysis (PCoAs) were performed to identify genetic clusters using the ade4 package (Dray & Dufour 2007) of R 3.2.1 (R development core team 2015). S7 coefficient of Gower and Legendre was used as a genetic distance since it gives more weight to shared insertions. Shared absences were not used because they do not give information about the genetic distance between individuals. Pairwise populations F_{ST} were computed using Arlequin 3.5 (Excoffier & Lischer 2010); significance of the index was assessed over 1,000 permutations using a significance threshold of 0.05.

The genomic scan was performed in two steps for each of the sampling replicates of each TE. First, Bayescan 2.1 (Foll & Gaggiotti 2008) was used to test for each locus deviation from neutrality. Bayescan consider a fission/island model where all subpopulations derive from a unique ancestral population. In this model, variance in allele frequencies between subpopulations is expected to be due either to the genetic drift that occurred independently in each subpopulation or to selection that is a locus-specific parameter. The differentiation at each locus in each subpopulation from the ancestral population is thus decomposed into a β component (shared by all loci in a

237 subpopulation) and is related to genetic drift, and a *α* component (shared for a locus by
238 all subpopulations) due to selection. Using a Bayesian framework, Bayescan tests for
239 each locus the significance of the *α* component. Rejection of the neutral model at one
240 locus is done using posterior Bayesian probabilities and controlled for multiple testing
241 using false discovery rate. In addition, Bayescan manage uncertainty about allele
242 frequency from dominant data such as the TD polymorphism, leaving the F_{IS} to freely
243 vary during the estimation of parameters. Bayescan was used with default values except
244 for the prior odds that were set to 100 (more compatible with datasets with a large
245 number of loci, see Bayescan manual), and a significance *q*-value threshold of 0.05 was
246 used to retain outlier loci. In a second step, only outliers suggesting divergent directional
247 selection between, Europe and Vietnam were considered. To identify them, locus-by-
248 locus Analyses of Molecular Variance (AMOVAs) were performed using Arlequin 3.5 for
249 each TE family. Significance of the F_{CT} (inter group differentiation) between Vietnamese
250 and European populations was assessed performing 10,000 permutations with a
251 significance threshold of 0.05. For each dataset, Bayescan outliers were crossed with
252 significant F_{CT} loci to retain candidate loci.
253 To identify the genomic environment of the candidate loci, the outlier sequences
254 (reference R2 read) were mapped onto the assembled genome of *Ae. albopictus* (Chen
255 *et al.* 2015) using Blastn. Blastn alignments were performed with default parameters and
256 sorted according to alignment score and after visual inspection of each alignment.
257 Outlier loci with multiple identical hits were discarded. To identify genes surrounding the
258 mapped outliers, the complete transcriptome of *Ae. albopictus* (including eggs, larvae
259 and adult females, downloaded at www.albopictusexpression.org Armbruster *et al.*) was

mapped over the reference genome using blat with default parameters; after alignment, one best hit was retained per transcript according to the best alignment score. When a transcript had multiple best hits, all positions for the transcript were considered.

PCR validation and Outlier analyses.

Pairs of primers were designed for each outlier locus in order to be used in standardized conditions. Forward primer was located in the TE end of the concerned family and reverse primer was set from the outlier locus (pairs of primers which successfully amplified insertions are provided in Table S2). All primer pairs were first tested on a set of 10 individuals in order to assess their specificity using 1/50 dilution of starting DNA from the TD experiment. Validated primers were then used to check the insertions polymorphism in 47 representative individuals from the 8 populations studied in the TD experiment using 1/50 dilutions of the starting DNA (not all individuals could be used because of DNA limitations). All PCRs were conducted in a final volume of 25μL using 0.5μL of diluted DNA, 0.5μL of each primer (10μM), 1μl of dNTPs (10mM) and 1U of DreamTaq Polymerase with 1X green buffer (ThermoFisher Scientific). Amplification was performed as follows: denaturation at 94C for 2 minutes then 34 cycles including denaturation at 94C for 30 seconds, hybridization at 60C for 45 seconds and elongation at 72C for 45 seconds; final elongation was performed for 10 minutes at 72C. After 45 minutes migration of the PCR product on 1X electrophoresis agarose gel, CG and MB assessed insertion polymorphism independently.

Results

High throughput TE insertion genotyping.

Sequencing produced a total of 102,319,300 paired-end reads (2x101bp). After quality and specificity filtering 24,332,715 reads were suitable for analyses. On average, a total number of 128,491 polymorphic insertion loci were available for each of the three sampling replicates. The mean number of loci per individual and per TE family ranged from 1025 ± 290 s.d. (IL1 family, mean and s.d. averaged over the three replicates) to 3266 ± 766 s.d. (RTE5 family). Details are given in Table S1. While our read sampling procedure could have artificially lowered the mean insertion frequency of the loci, this effect should be small because in our final datasets, the TE insertion frequencies (*i. e.* the number of individuals that share an insertion) are not correlated with the mean number of read per individual at the considered locus (Figure S3).

Population structure.

Principal Coordinate Analyses (PcoAs) were performed independently for each of the five TEs (Figure 2). Among the three main Principal Coordinates (PCs), individuals tend to be grouped according to their respective populations with little overlap between groups. However, the three main PCs represent only a small fraction of the total variance (< 10%), suggesting a weak genetic structuring between the populations. Overall, individuals from Vietnamese populations (HCM, TA, VT) tend to be grouped

301 together in a single cluster, at the exception of 13 to 14 individuals from HCM when
302 using L2B and RTE5 TE families (S2 Figure), along with six individuals of VT with the
303 RTE4 TE family (Figure 1) that can not be clearly distinguished from European samples.
304 BCN individuals (Spain) represent the most homogeneous group, well differentiated
305 from Vietnamese and French individuals (SP, CGN, NCE and PLV).
306 In agreement with PCoAs, Analyses of Molecular Variance (AMOVAs) attributed very
307 few genetic variances among groups (Vietnam-Europe) and between populations within
308 groups (Table 1). In the studied populations, most of the genetic variance was
309 distributed among individuals within groups.

310 Measures of genetic differentiation among pairs of populations were consistent with
311 PCoAs and AMOVAs (File S1): the BCN population shows the highest F_{ST} values with
312 the other populations for each of the five TEs ($0.051 < F_{ST} < 0.148$), while Vietnamese
313 populations were the most closely related ($0.011 < F_{ST} < 0.032$). While VT is located
314 100 km away from TA and HCM (both sampled in the same city, Hô Chi Minh, Vietnam)
315 the F_{ST} values are very similar between the three Vietnamese populations, suggesting
316 no influence of geography at this scale. CGN and NCE, sampled in the same urban area
317 (Nice agglomeration), are also little or not significantly differentiated depending on the
318 TE family. The previously identified intermediate pattern of HCM with some European
319 populations at L2B and RTE5 loci (PCoAs analyses) is also found at the F_{ST} level,
320 especially regarding the low differentiation with the PLV population for these markers
321 ($0.011 < F_{ST} < 0.020$).

322

323 **Genomic scan.**

324 Research of outlier loci for both selection signature using Bayescan (non-hierarchical
325 island model), and for significant F_{CT} (between Europe-Vietnam group differentiation)
326 identified 92 candidate insertion loci (Figure 3). Most of these insertions are found in
327 both areas (no private allele), except for RTE4_6 and RTE4_7 that were not found in
328 Vietnam. In addition, a majority of outliers corresponds to high frequency insertions in
329 Europe, while the same trend is not observed at 92 randomly chosen loci among those
330 having the same minimum insertion frequency (≥ 20 individuals/locus) between Europe
331 and Vietnam (Figure 4). PCR amplification of the outlier loci were carried out on a
332 representative panel of 47 individuals to validate the insertion pattern detected by TD
333 (see Material and Method). For loci where the amplification was successful, the insertion
334 pattern observed by PCR always confirmed the results from TD (Figure S5).

335 From 92 outlier loci, 21 could be attributed to a unique position on the *Ae. albopictus*
336 genome. Annotation and distance to surrounding genes are reported on S2 File. We
337 found that six outliers (S2 File, sheet 3, highlighted) loci are located on contigs that
338 harbor genes previously identified in *Ae. albopictus* as being differentially expressed
339 between diapause-induced and non diapause-induced samples , and for which
340 orthologs in *Drosophila melanogaster* are known to be part of well-identified functional
341 networks (Poelchau *et al.* 2013b). All these six loci are found to be outliers because of
342 their high insertion frequencies in Europe compared with Vietnam.

Discussion

The goal of our study was to identify genomic regions involved in adaptive evolution of *Ae. albopictus* thanks to the development of new genetic markers. Through high-throughput genotyping of five TE families' insertion polymorphisms, we identified up to 128,617 polymorphic loci among a hundred of individuals from eight sampling sites. The estimated genome size of *Ae. albopictus* exceeds one billion base-pairs (Goubert *et al.* 2015; Dritsou *et al.* 2015; Chen *et al.* 2015). Accordingly, the amount of markers scored in this study offers a comfortable genomic density of one marker every 10 kb.

We provide here a new and cost efficient method to quickly generate a large amount of polymorphic markers without extensive knowledge about one species genome. Specifically, this strategy appears extremely valuable for species with a large genome size, where TE density could severely compromise the development of more classical approaches, such as the very popular RAD-sequencing.

The genetic structure of the studied populations showed strong consistency between sampling replicates of individuals' reads, demonstrating the robustness of the method in spite of an initial substantial coverage variation among individuals. Population genetics analyses revealed a very low level of genetic structuring between European and Vietnamese populations. Among the studied populations, AMOVAs showed that most of the genetic variation is distributed between individuals within populations (> 90%), and as suggested by pairwise F_{ST} and PCoAs, only a small part (< 10%) of the genetic variance is due to differentiation between populations. The genetic differentiation we

364 measured is indeed as high among European populations as it is between populations
365 from Europe and Vietnam.

366 This singular population structure is in agreement with previous results gathered in *Ae.*
367 *albopictus* using different collections of allozymes, mtDNA or microsatellites markers
368 (Black *et al.* 1988; Kambhampati *et al.* 1991; Zhong *et al.* 2013; Gupta & Preet 2014;
369 Manni *et al.* 2015). Moreover, a recent analysis performed with a set of 11
370 microsatellites on individuals from the same populations (at the exception of BCN)
371 showed a similar distribution of genetic variation among hierarchical levels (Minard *et al.*
372 2015). These results demonstrate the reliability of our markers and confirm that a non-
373 hierarchical island model can likely fit the global genetic structure. The genetic diversity
374 observed in Europe is compatible with a scenario of multiple and independent
375 introductions, as already suggested for *Ae. albopictus* (Urbanelli *et al.* 2000; Birungi &
376 Munstermann 2002; Takumi *et al.* 2009; Becker *et al.* 2013). However, as previously
377 suggested, this pattern could also be the result of founder events that may occur during
378 colonization and/or a restriction of gene flow between populations consecutive to their
379 introduction. Answering such a question would require an extended sampling all over
380 the native area.

381 Outlier analysis revealed 92 loci with high posterior probabilities of being under positive
382 selection between European and Vietnamese populations. When possible, the PCR
383 amplification of the outlier loci, using a set of representative individuals, always
384 confirmed a shift of insertion frequencies toward either the European or the Vietnamese
385 sampling sites. This suggests that in spite of a reduced coverage, introduced by

386 sampling in the dataset, the scored insertion polymorphisms are reliable. In addition, our
387 method of analysis is likely to be conservative: the Bayescan outliers were selected for
388 their consistency with a significant F_{CT} between European temperate and Vietnamese
389 tropical populations, which avoid retaining outliers that we were not looking for, for
390 example those due to a population specific event.

391 We were able to assign a unique position for 21 of the outlier loci on the *Ae. albopictus*
392 genome. As expected by the *Ae. albopictus* genomic composition (Goubert *et al.* 2015;
393 Dritsou *et al.* 2015; Chen *et al.* 2015), an important part of the other outlier loci were
394 located in repeated regions (44,6% of total), despite our efforts to remove *a priori* loci
395 occurring in known transposable elements. Since the *Ae albopictus* genome publication
396 is very recent, no gene set or other genome annotation are currently available. We thus
397 took advantage of the *Ae. albopictus* transcriptome data (Poelchau *et al.* 2013a) to
398 annotate regions surrounding the detected ouliers. We found six outliers located in
399 contigs which also harbor genes differentially expressed between individuals induced for
400 diapause and controls (Poelchau *et al.* 2013b), two of them, being located either in an
401 intron (RTE4_7442) or within 3kb (RTE4_17015) of these candidate genes. It is worth
402 mentioning that these two genes belong to the same functional group (GO:0005576
403 extracellular region). Diapause is a critical developmental stage found only in temperate
404 populations of the Asian tiger mosquito. Interestingly, this functional pathway has been
405 shown to benefit from fast adjustments thanks to local adaptation. For instance,
406 Urbanski *et al.* (2012) showed that invasive American populations originating from
407 Japan have rapidly evolved a new adaptive clinal response to diapause induction,

408 independent from that observed in the native area. Thus, adaption in the temperate
409 regions could have led to several selective sweeps on gene or regulatory sequences
410 involved in this critical pathway, allowing the settlement of the mosquito in new
411 temperate areas.

412 Interestingly, and as it is the case for these six outliers, we found significantly more
413 outlier loci with a high frequency in Europe and low frequency in Vietnam than the
414 opposite pattern. This was unexpected regarding our initial assumptions: a favored allele
415 selected in one or another environment has *a priori* no reason to be more often
416 associated with the presence or the absence of a TE insertion at linked sites. However,
417 we found that the majority of the sequenced TE insertions segregates at low frequencies
418 (around 10% of all individuals). When considering the linked region of one polymorphic
419 TE insertion, if a favorable mutation appears in an individual where the insertion is
420 absent, the increase of frequency of this “absence” haplotype will thus, most of the time,
421 have a modest effect on the genetic differentiation at this marker, since it is already
422 segregating at high frequency. By contrast, if a favorable mutation appears in a TE
423 “presence” haplotype, the increase in frequency of the linked TE insertion would lead to
424 high differentiation (F_{CT}) values. In absence of an alternative explanation, our outlier loci
425 could thus indicate in which subset of populations the adaptive mutation occurred, and
426 in the present case, this would have happened more frequently in the temperate
427 populations.

428 Two scenarios, not mutually exclusive, could be invoked in the light of our data. A simple
429 case would be a direct adaptive evolution in European invasive population that

originated from tropical regions of the native area. A second hypothesis, could be that invasive temperate populations came from Northernmost territories of the native area such as northern China or Japan where *Ae. albopictus* populations are already cold-adapted. It would be thus interesting to know whether the observed signature of selection results from more "ancient" adaptations in the native area, or if it originates from more recent fine tuning of cold-related traits in the invasive areas. A recent study (Porretta *et al.* 2012) suggested, using new variable *COI* mtDNA sequences and historical species range modeling, that Northern areas of the native range of *Ae. albopictus* would be the latest to have been colonized after a range expansion from Southern refugia following the last glacial around 21,000 years ago. The authors suggested that *Ae. albopictus* may have followed the human populations during their expansion from South to North in this area, that began approximately 15,000 years ago. Thus wherever the origin of the invasive individuals sampled in Europe, it is likely that they are representatives of populations that had recently undergone a shift of selective pressure from tropical to temperate climatic conditions. This could explain why so many outliers are associated with high insertion frequency in Europe, and that candidate genes in the diapause pathway are found in the neighborhood of some of these outliers. An easy way to distinguish between these possibilities would be to search if the same outlier insertions are present in several temperate populations from the native area.

It is important to note that the results presented here are only restricted to a subset of the Asian tiger mosquito populations located in temperate and tropical environments. It is thus probable that some of the outliers detected could be specific to this particular

comparison and do not reflect the global pattern of differentiation between tropical and temperate populations. Research of the same outliers between other tropical and temperate populations from the native and non-native areas would be extremely valuable to extrapolate our results at a larger scale. Should the same outlier insertions be found at high frequencies in temperate locations – such as in USA, Japan or China –, extended investigations about the origin of invasive populations would help clarify if those similarities are due to an ancestral sweep or parallel sweeps that occurred independently in several populations. This study already provides for some candidate loci a set of functional primers that could be directly used to answer this question in any DNA sample of *Ae. albopictus*.

We report here the first leads supporting adaptive evolution at the molecular level in the Asian tiger mosquito. Progress in the annotation of published genomes, and the looming availability of supplementary genomic resources will allow gaining the most from these results. We hope that this work will contribute to unravel the implication of adaptive processes during the invasion of disease vectors.

Acknowledgements

We thank Van Tran-Van, Christophe Bellet, Grégory Lambert, Huynh Kim Ly Khanh and Trang Huynh who made possible and contributed to the samplings in France and Vietnam. We are grateful to Valèria Romero Soriano and her family for their help during sampling in Sant Cugat dèl Vallès. Library construction and sequencing was made in

collaboration with Clémence Genthon and Olivier Bouchez. We thank Manon Vigneron for PCR validation experiments. We also thank Rita Rebollo who provided insightful comments and English revision of the manuscript. This work was performed using the computing facilities of the CC LBBE/PRABI. C.G. received a grant from the French Ministry of Superior Education. This work was supported by the Centre National de la Recherche Scientifique, the Institut Universitaire de France, and preliminary experiments benefited from of grant of the Federation de Recherche 41 “Bio-Environnement et Santé”. ”. Funding for mosquito sampling in Vietnam was provided by grants from EC2CO CNRS and occurred within the framework of GDRI “Biodiversity and Infectious Diseases in Southeast Asia”. Original maps used to describe sampling were freely downloaded from http://d-maps.com/carte.php?num_car=4719&lang=en and http://d-maps.com/carte.php?num_car=708&lang=en

References

Akkouche A, Rebollo R, Burlet N *et al.* (2012) tirant, a newly discovered active endogenous retrovirus in Drosophila simulans. *Journal of virology*, **86**, 3675–81.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–10.

Becker N, Geier M, Balczun C *et al.* (2013) Repeated introduction of Aedes albopictus into Germany, July to October 2012. *Parasitology research*, **112**, 1787–90.

Biedler J, Qi Y, Holligan D *et al.* (2003) Transposable element (TE) display and rapid detection of TE insertion polymorphism in the Anopheles gambiae species complex. *Insect Molecular Biology*, **12**, 211–216.

Birungi J, Munstermann LE (2002) Genetic Structure of Aedes albopictus (Diptera: Culicidae) Populations Based on Mitochondrial ND5 Sequences: Evidence for an

499 Independent Invasion into Brazil and United States. *Annals of the Entomological*
500 *Society of America*, **95**, 125–132.

501 Black WC, Ferrari JA, Sprengert D (1988) Breeding structure of a colonising species :
502 Aedes albopictus (Skuse) in the United States. , **60**.

503 Bock DG, Caseys C, Cousens RD *et al.* (2015) What we still don't know about invasion
504 genetics. *Molecular Ecology*, n/a–n/a.

505 Bonin A, Paris M, Després L *et al.* (2008) A MITE-based genotyping method to reveal
506 hundreds of DNA polymorphisms in an animal genome after a few generations of
507 artificial selection. *BMC genomics*, **9**, 459.

508 Bonin A, Paris M, Tetreau G, David J-P, Després L (2009) Candidate genes revealed by
509 a genome scan for mosquito resistance to a bacterial insecticide: sequence and
510 gene expression variations. *BMC genomics*, **10**, 551.

511 Bonizzoni M, Gasperi G, Chen X, James AA (2013) The invasive mosquito species
512 Aedes albopictus: current knowledge and future perspectives. *Trends in*
513 *parasitology*, **29**, 460–468.

514 Boulesteix M, Simard F, Antonio-Nkondjio C *et al.* (2007) Insertion polymorphism of
515 transposable elements and population structure of Anopheles gambiae M and S
516 molecular forms in Cameroon. *Molecular ecology*, **16**, 441–452.

517 Carnelossi EAG, Lerat E, Henri H *et al.* (2014) Specific activation of an I-like element in
518 Drosophila interspecific hybrids. *Genome biology and evolution*, **6**, 1806–17.

519 Casacuberta E, González J (2013) The impact of transposable elements in
520 environmental adaptation. *Molecular ecology*, 1503–1517.

521 Chen X-G, Jiang X, Gu J *et al.* (2015) Genome sequence of the Asian Tiger mosquito,
522 Aedes albopictus, reveals insights into its biology, genetics, and evolution.
523 *Proceedings of the National Academy of Sciences of the United States of America*,
524 **112**, E5907–5915.

525 Chouin-Carneiro T, Vega-Rua A, Vazeille M *et al.* (2016) Differential Susceptibilities of
526 Aedes aegypti and Aedes albopictus from the Americas to Zika Virus (MJ Turell,
527 Ed,). *PLOS Neglected Tropical Diseases*, **10**, e0004543.

528 Colautti RI, Lau JA (2015) Contemporary evolution during invasion: evidence for
529 differentiation, natural selection, and local adaptation. *Molecular Ecology*, **24**, 1999–
530 2017.

531 Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2012) Special features of RAD
532 Sequencing data: implications for genotyping. *Molecular ecology*.

533 Dray S, Dufour A-B (2007) The ade4 Package: Implementing the Duality Diagram for
534 Ecologists. *Journal of Statistical Software*, **22**, 1–20.

535 Dritsou V, Topalis P, Windbichler N *et al.* (2015) A draft genome sequence of an
536 invasive mosquito: an Italian Aedes albopictus. *Pathogens and global health*,
537 2047773215Y0000000031.

538 Esnault C, Boulesteix M, Duchemin JB *et al.* (2008) High genetic differentiation between
539 the M and S molecular forms of Anopheles gambiae in Africa. (P Awadalla, Ed,).
540 *PloS one*, **3**, e1968.

541 Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to
542 perform population genetics analyses under Linux and Windows. *Molecular ecology*
543 *resources*, **10**, 564–7.

544 Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate
545 for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**,
546 977–93.

547 Goubert C, Modolo L, Vieira C *et al.* (2015) De novo assembly and annotation of the
548 Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw
549 genomic reads and comparative analysis with the yellow fever mosquito (Aedes
550 aegypti). *Genome Biology and Evolution*, evv050–.

551 Goubert C, Minard G, Vieira C, Boulesteix M (2016) Population genetics of the Asian
552 tiger mosquito *Aedes albopictus*, an invasive vector of human diseases. *Heredity*,
553 10.1038/hdy.2016.35.

554 Grard G, Caron M, Mombo IM *et al.* (2014) Zika Virus in Gabon (Central Africa) – 2007:
555 A New Threat from Aedes albopictus? (R Charrel, Ed,). *PLoS Neglected Tropical*
556 *Diseases*, **8**, e2681.

557 Gupta S, Preet S (2014) Genetic differentiation of invasive Aedes albopictus by RAPD-
558 PCR: Implications for effective vector control. *Parasitology Research*, **113**, 2137–
559 2142.

560 Handley L-J, Estoup A, Evans DM *et al.* (2011) Ecological genetics of invasive alien
561 species. *BioControl*, **56**, 409–428.

562 Hanson SM, Craig GB (1994) Cold Acclimation, Diapause, and Geographic Origin Affect
563 Cold Hardiness in Eggs of Aedes albopictus (Diptera: Culicidae). *Journal of Medical*
564 *Entomology*, **31**, 192–201.

565 Hawley WA (1988) The biology of Aedes albopictus. *Journal of the American Mosquito*
566 *Control Association. Supplement*, **1**, 1–39.

567 Hawley W, Reiter P, Copeland R, Pumpuni C, Craig G (1987) Aedes albopictus in North
568 America: probable introduction in used tires from northern Asia. *Science*, **236**,
569 1114–1116.

570 Kambhampati S, Black WC, Rai KS (1991) Geographic origin of the US and Brazilian
571 Aedes albopictus inferred from allozyme analysis. *Heredity*, **67 (Pt 1)**, 85–93.

572 Lande R (2015) Evolution of phenotypic plasticity in colonizing species. *Molecular*
573 *ecology*.

574 Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of
575 protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, **22**, 1658–9.

576 Manni M, Gomulski LM, Aketarawong N *et al.* (2015) Molecular markers for analyses of
577 intraspecific genetic diversity in the Asian Tiger mosquito, Aedes albopictus.
578 *Parasites & vectors*, **8**, 188.

579 Marcondes CB, Ximenes M de FF de M (2015) Zika virus in Brazil and the danger of
580 infestation by Aedes (Stegomyia) mosquitoes. *Revista da Sociedade Brasileira de*
581 *Medicina Tropical*.

582 Martin M (2011) Cutadapt removes adapter sequences from high-throughput
583 sequencing reads. *EMBnet.journal*, **17**, 10.

584 Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-
585 effective polymorphism identification and genotyping using restriction site
586 associated DNA (RAD) markers. *Genome research*, **17**, 240–8.

587 Minard G, Tran F-H, Tran-van V *et al.* (2015) French invasive Asian tiger mosquito
588 populations harbor reduced bacterial microbiota and genetic diversity compared to
589 Vietnamese autochthonous relatives. *Frontiers in Microbiology*, **6**.

590 Modolo L, Lerat E (2015) UrQt: an efficient software for the Unsupervised Quality
591 trimming of NGS data. *BMC Bioinformatics*, **16**, 137.

592 Monden Y, Yamamoto A, Shindo A, Tahara M (2014) Efficient DNA fingerprinting based
593 on the targeted sequencing of active retrotransposon insertion sites using a bench-
594 top high-throughput sequencing platform. *DNA research : an international journal for*
595 *rapid publication of reports on genes and genomes*, **21**, 491–8.

596 Munroe DJ, Haas M, Bric E *et al.* (1994) IRE-bubble PCR: a rapid method for efficient
597 and representative amplification of human genomic DNA sequences from complex
598 sources. *Genomics*, **19**, 506–14.

599 Paupy C, Delatte H, Bagny L, Corbel V, Fontenille D (2009) Aedes albopictus, an
600 arbovirus vector: from the darkness to the light. *Microbes and infection / Institut*
601 *Pasteur*, **11**, 1177–85.

602 Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of
603 standing genetic variation. *Molecular ecology*.

604 Poelchau MF, Reynolds J a., Denlinger DL, Elsik CG, Armbruster P a. (2013a)
605 Transcriptome sequencing as a platform to elucidate molecular components of the
606 diapause response in the Asian tiger mosquito, Aedes albopictus. *Physiological*
607 *entomology*, **38**, 173–181.

Poelchau MF, Reynolds J a, Elsik CG, Denlinger DL, Armbruster P a (2013b) Deep sequencing reveals complex mechanisms of diapause preparation in the invasive mosquito, Aedes albopictus. *Proceedings. Biological sciences / The Royal Society*, **280**, 20130143.

Porretta D, Mastrantonio V, Bellini R, Somboon P, Urbanelli S (2012) Glacial history of a modern invader: phylogeography and species distribution modelling of the Asian tiger mosquito Aedes albopictus. *PloS one*, **7**, e44515.

Roy AM, Carroll ML, Kass DH *et al.* (1999) Recently integrated human Alu repeats: finding needles in the haystack. *Genetica*, **107**, 149–161.

Santolamazza F, Mancini E, Simard F *et al.* (2008) Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms. *Malaria journal*, **7**, 163.

Stapley J, Santure AW, Dennis SR (2015) Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Molecular ecology*, **24**, 2241–52.

Takumi K, Scholte E-J, Braks M *et al.* (2009) Introduction, scenarios for establishment and seasonal activity of Aedes albopictus in The Netherlands. *Vector borne and zoonotic diseases (Larchmont, N.Y.)*, **9**, 191–6.

Urbanelli S, Bellini R, Carrieri M, Sallicandro P, Celli G (2000) Population structure of Aedes albopictus (Skuse): the mosquito which is colonizing Mediterranean countries. *Heredity*, **84 (Pt 3)**, 331–337.

Urbanski JM, Benoit JB, Michaud MR, Denlinger DL, Armbruster P (2010) The molecular physiology of increased egg desiccation resistance during diapause in the invasive mosquito, Aedes albopictus. *Proceedings. Biological sciences / The Royal Society*, **277**, 2683–92.

Urbanski J, Mogi M, O'Donnell D *et al.* (2012) Rapid adaptive evolution of photoperiodic response during invasion and range expansion across a climatic gradient. *The American naturalist*, **179**, 490–500.

Zhong D, Lo E, Hu R *et al.* (2013) Genetic analysis of invasive Aedes albopictus populations in Los Angeles County, California and its potential public health impact. *PloS one*, **8**, e68586.

Data Accessibility

Paired-end raw sequences are available through SRA at NCBI under SRP070185

642 (Bioproject PRJNA312147)

643 Final presence/absence matrices (including replicates) are available at Dryad
644 (doi:10.5061/dryad.9p925) at http://datadryad.org/review?doi=doi:10.5061/dryad.9p925

645 Author contributions

646 CG, CV and MB conceived the experiments and conducted the analyses. CG and HH
647 developed and performed the molecular experiments. GM, CVM and PM conducted the
648 sampling in France and Vietnam. All authors contributed to the final version of the
649 manuscript.

Tables and Figures

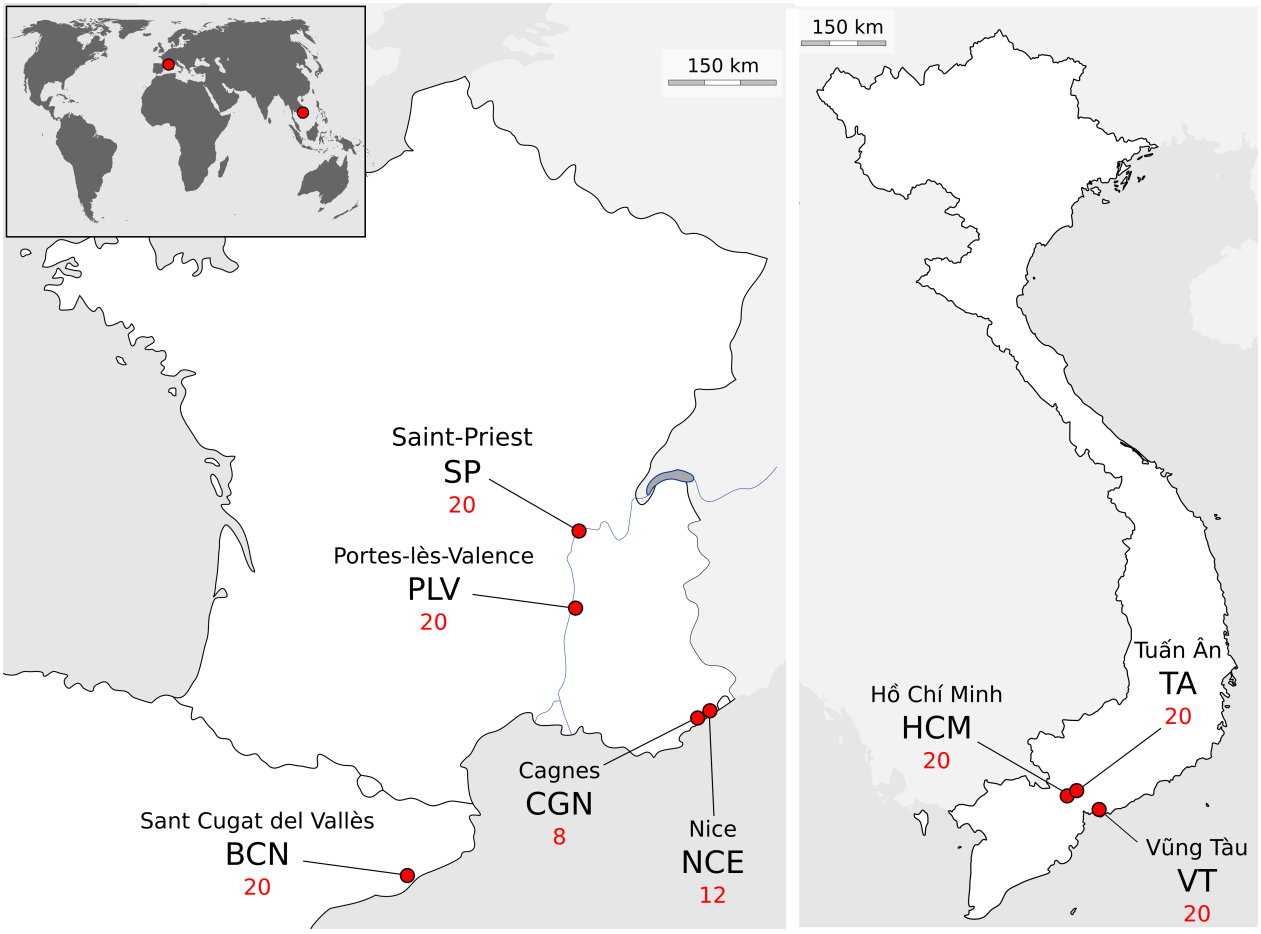


Figure 1. Sampling sites (with abbreviations) of *Ae. albopictus* in Europe and Vietnam. Red numbers correspond to the total number of individual sampled. Supporting information about samples are available in Table S1.

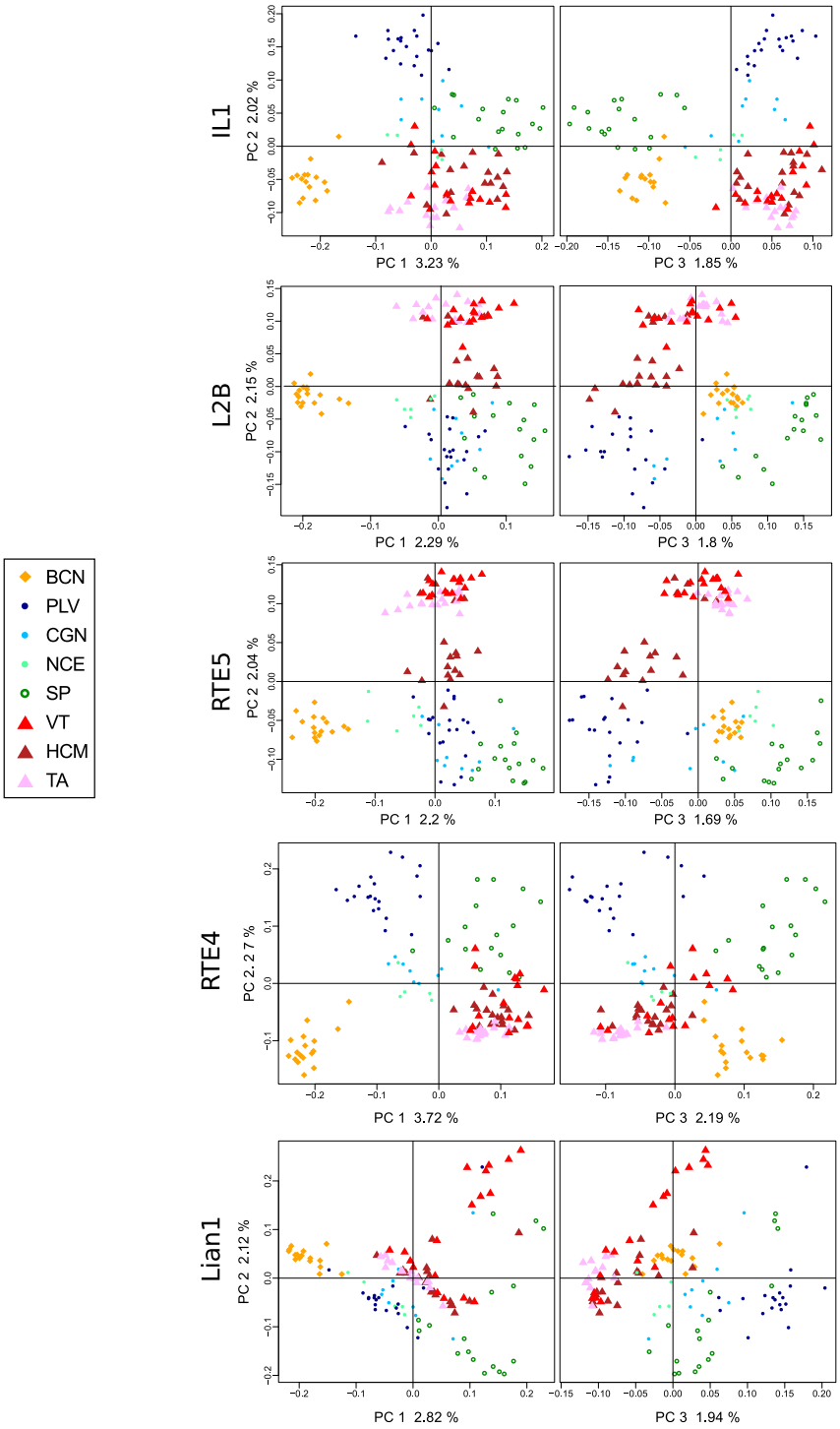


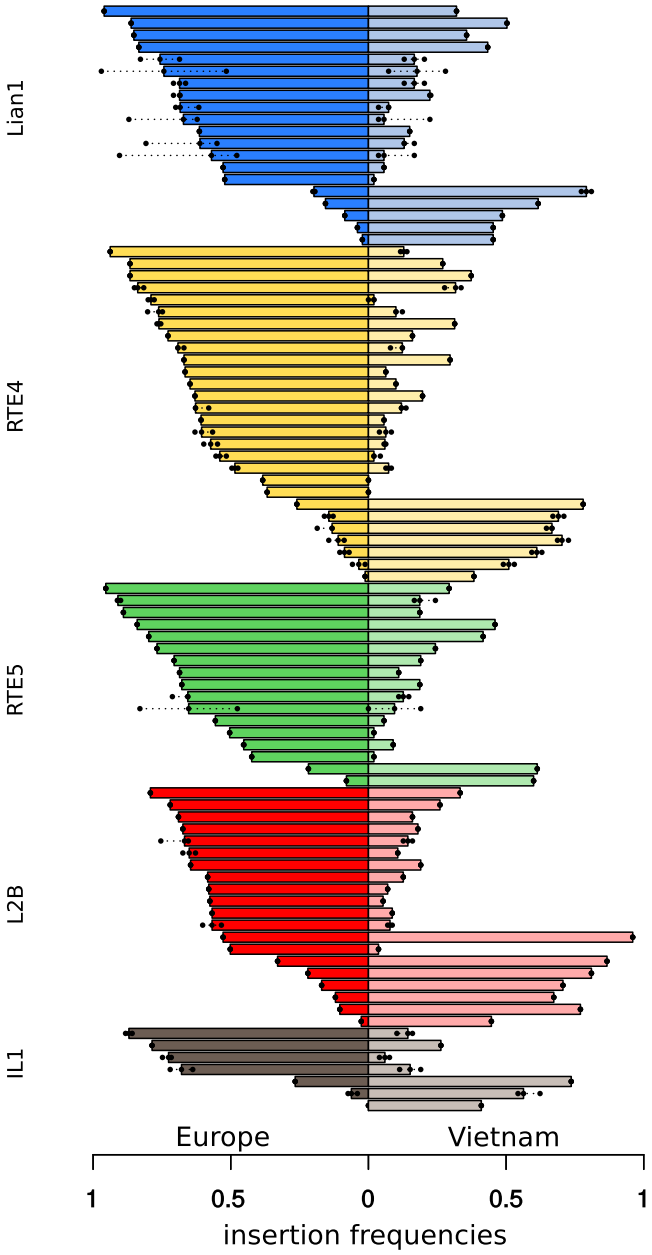
Figure 2. Principal Coordinate Analyses (PCoAs). Projection of individuals over the three first principal coordinates (PC) of PCoAs for each of the 5 TE families and for the first replicate (M1, see Material and Methods). Proportion of inertia represented by each axes is noted in %. circles: European populations; triangles: Vietnamese populations. Results for other sampling replicate can be found in Figure S4.

661 **Table 1. Analyses of Molecular Variance (AMOVAs).** Results for the three replicates
662 (M1, M2, M3) of read sampling for the five TE families (IL1, L2B, RTE5, RTE4, Lian1).
663 Values are given in percentage of the total genetic variance

	IL1	L2B	RTE5	RTE4	Lian1
Among groups	[0.59-0.70]	[1.22-1.29]	[1.08-1.10]	[1.97-2.04]	[0.67-0.74]
Among populations within groups	[5.15-5.37]	[3.58-3.63]	[3.36-3.40]	[6.67-6.78]	[4.47-4.56]
Within populations	[94.04-94.16]	[95.08-95.18]	[95.51-95.55]	[91.18-91.30]	[94.77-94.81]

intervals reports min and max values among the 3 sampling replicates

664



665

666 **Figure 3. Insertion frequencies in Europe and Vietnam for the 92 outlier loci**. Bars
667 represent the median value from the three reads sampling replicates and dots the
668 values from the other replicates (if outlier found in replicate). Colors correspond to each
669 of the 5 TE families.

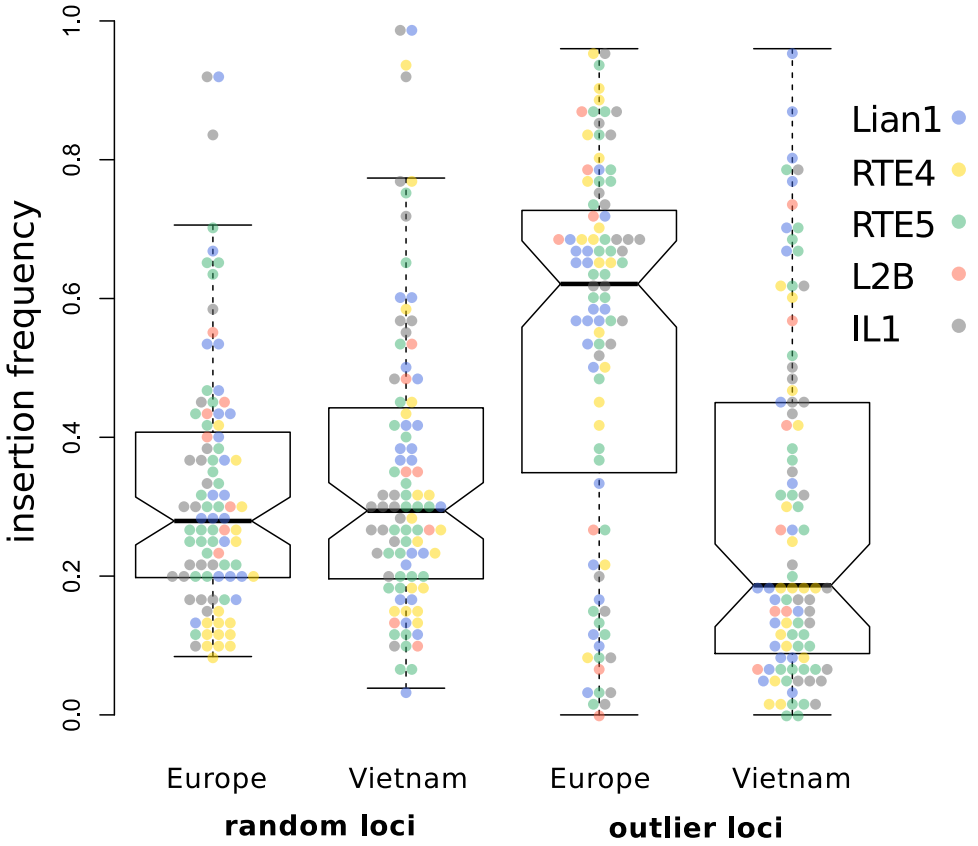


Figure 4. Comparison of outliers frequencies with randomly selected loci. Insertion frequencies of 92 randomly chosen loci among those having the same minimum insertion frequency (≥20 individuals) as outliers compared to the 92 outlier loci. Random loci were taken from the first replicate (M1) and values for outliers are median values obtained among the three replicates. Non-overlapping notches indicate a significant difference between the true medians (thick dark horizontal bars).